Int Tax Public Finance
DOI 10.1007/s10797-014-9304-5

Empirical welfare analysis with preference heterogeneity

André M. J. Decoster · Peter Haan

Received: 10 May 2013 / Accepted: 22 January 2014

Abstract We apply recently proposed individual welfare measures in the context of preference heterogeneity, derived from structural labour supply models. Contrary to the standard practice of using reference preferences and wages, these measures preserve preference heterogeneity in the normative step of the analysis. They also make the ethical priors, implicit in any interpersonal comparison, more explicit. Information on preference heterogeneity is obtained from a structural discrete choice labour supply model for married women estimated on microdata from the Socio Economic Panel in Germany. We construct welfare orderings of households according to the different metrics, each embodying different ethical choices concerning the treatment of preference heterogeneity in the consumption-leisure space and provide empirical evidence about the sensitivity of the welfare orderings to different normative principles. We also discuss how sensitive the assessment of a tax reform is to the choice of different metrics.

Keywords Welfare measures · Labour supply · Preference heterogeneity

We are grateful to participants of the seminars "Basic Income Policies and Tax Reforms: Alternative Concepts and Evaluations" at the University of Turin, 19–20 November 2009, "Fairness, Equality of Opportunity and Public Economics" at CORE, Louvain-la-Neuve, 9–10 April 2010 and the EUROMOD workshop at ISER (University of Essex) 23–24 September 2010, for useful comments on the results presented in this paper. Peter Haan gratefully acknowledges financial support by the Thyssen Foundation in Project AZ.10.11.2.085. The usual disclaimer applies.

A. M. J. Decoster (✉)
Department of Economics, KU Leuven, Louvain, Belgium
e-mail: andre.decoster@kuleuven.be

P. Haan
DIW, Berlin, Germany
e-mail: phaan@diw.de

P. Haan
Freie Universität, Berlin, Germany

Published online: 19 February 2014

JEL Classification C35 · D63 · D78 · H24 · H31 · J22

1 Introduction

The aim of this paper is to provide empirical evidence that the choice of the normative framework used to evaluate policy reforms which affect the labour-leisure choice strongly influences the welfare analysis of the reform. More specifically, we investigate the role of preference heterogeneity in the normative step of the analysis. Obviously, preference heterogeneity plays a prominent role in the positive part of the analysis, where substantial progress has been made in modelling individual labour supply decisions based on the structural specification of preferences. The feasibility of these models to account for complicated real-world budget constraints and their ease of interpretation make them especially attractive for the ex ante evaluation of policy reforms in the tax benefit sphere. For an overview, see Creedy and Kalb (2005).

The quick dissemination of these models makes it all the more surprising that, at least until recently and certainly in the applied literature, much less attention has been devoted to the normative implications of preference heterogeneity. One of the reasons might be that interpersonal welfare comparisons are non-trivial in a utilitarian framework in which individuals not only differ in abilities, but also in preferences. As Boadway (2012, p. 517) rightly remarks, by assuming that individuals only differ in abilities, but otherwise have identical preferences, this 'problem has largely been side-stepped in the mainstream normative second-best policy literature'.

In applied welfare analysis, however, certainly when it aims to provide policy makers with an overall welfare assessment of a policy change, this position is much more difficult to maintain. One wants to discriminate between low labour incomes coming, on the one hand, from low productivity (or innate ability), and, on the other, from a high preference for leisure. In classical applied welfare analysis, individual welfare metrics such as equivalent or compensating variations, both based on money metric utilities, are known well enough. But in a context of individuals with heterogeneous preferences, comparability and/or aggregation of these metrics faces serious difficulties. Simply stated, when indifference curves cross, the ordering of individuals in terms of better or worse off is easily reversed by choosing another reference price to calculate the money metric. And when prices are individual specific (such as wages), using this individual-specific wage in the money metric will assign a different welfare level to two individuals who have the same preferences and are on the same indifference curve (see Boadway and Bruce 1984, Chapter 9; or Auerbach 1985).

To deal with these problems of interpersonal comparability, one can identify two tracks in the relevant literature. Given the importance of preference heterogeneity for their positive models, a first bunch of authors explicitly acknowledges the challenge of using classical individual welfare metrics in their context of preference heterogeneity.[1] Yet, they then proceed by calculating individual money metrics and either aggregate

[1] See e.g. footnote 5 on p. 804 of Eissa et al. (2008), or Creedy and Hérault (2012) p. 131: "While the difficulties associated with metrics are recognized, it is also the case that, as Donaldson (1992, p. 89) stressed, 'no methodology in applied welfare economics is perfect"'.

them in an unweighted (Eissa et al. 2008) or weighted sum (Creedy and Hérault 2012). As such there is nothing wrong with calculating weighted sums of individual equivalent variations, and we certainly agree with the quote referred to by Creedy and Hérault (2012) in footnote 1 above. But the use and aggregation of this kind of welfare metrics introduces implicit comparability assumptions which would preferably have been made on an explicit basis. Only then is it possible to discuss the ethical priors underlying e.g. legitimate claims for compensation by worse-off individuals.

The second strand in the literature follows King (1983) who implements interpersonal comparability by evaluating chosen bundles by means of a reference preference ordering (the so-called reference household) at reference prices. Recent applications of this approach in the context of labour supply concern e.g. Aaberge et al. (2004) and Aaberge and Colombino (2013) To simulate labour supply responses to tax reforms, they estimate and use preferences which are heterogeneous across households. But when moving from the positive into the normative step of the analysis, actual preferences are replaced by a common preference ordering of a reference household. It is true that in this specific case, this common utility function is itself estimated on a sample of individuals with heterogeneous preferences. But this does not diminish the fact that (only) in the normative part of the analysis, preference heterogeneity itself is removed from the scene.[2] The normative literature on interpersonal comparisons has therefore christened this procedure as 'Perfectionism'.[3]

The latter term clearly reveals what is at stake, since in this normative literature this "Perfectionism" is opposed to another property of social orderings, viz. one which expresses in one form or another "respecting preferences" of the individuals. In Sect. 2, we will briefly summarize how the attempt to respect preferences of individuals in the construction of the social ordering (mostly called Paretianity) clashes with even some weak and intuitive forms of making interpersonal comparisons (such as bundle dominance, which we will use as an example in Sect. 2). The literature mentioned in the previous paragraph escapes the clash by removing preference heterogeneity and imposing preferences determined by the social planner. Yet, precisely the research into this clash between forms of interpersonal comparability and Paretianity (or respecting individual preferences) in a context of preference heterogeneity has proven to be fruitful to discover new and complementary perspectives in designing individual welfare metrics in heterogeneous environments. This research, summarized in Fleurbaey (2008) and Fleurbaey and Maniquet (2011), shows how to construct a normative framework which maximally retains preference heterogeneity, and how individual welfare metrics follow from this analysis.

In this paper, we demonstrate the usefulness of these individual welfare metrics in the context of empirically estimated heterogeneous preferences. The key feature of

[2] Contrary to what is often thought, a sensitivity analysis does not introduce genuine preference heterogeneity into the normative analysis. In each step of the sensitivity analysis, *all* individuals or households are endowed with the same preference ordering.

[3] The point is not that the normative tool becomes disconnected from the positive tool(s), since this is the essence of a normative position (see e.g. Creedy and Hérault 2011 or Capéau et al. 2009 for interpretations and applications, based on a similar explicit distinction between the positive and normative step, and Manski (2012) for an explicit plea to distinguish both steps and a critical position as far as current knowledge of consumption-leisure preferences is concerned to inform tax policy).

the metrics introduced below is that they fully respect preferences: all metrics increase when the individual moves to a bundle on a higher indifference curve of her own preference ordering. But we also illustrate one of the major advantages of these individual welfare metrics, to wit that they bring the normative choices clearer to the surface. Indeed, once we allow for heterogenous preferences across individuals, the non-trivial issue of making well-founded interpersonal comparisons of well-being re-enters the scene in an ethically even richer way. If one removes preference heterogeneity from the normative analysis, people only differ in abilities and non-labour income. But with preference heterogeneity preserved, one also needs a fairness concept which takes into account that individual outcomes not only result from endowed circumstances, but also from individual preferences. The individual welfare metrics used in this paper embody different ethical priors on how to treat preference heterogeneity, which, in the context of differences in willingness to work, boils down to either favouring the industrious or the work averse. We demonstrate in a highly relevant empirical context of preferences between consumption and leisure, how the underlying ethical choices systematically alter interpersonal comparisons of well-being.

In this respect, our paper can be read as a complement to Preston and Walker (1999). These authors lined up many of the measures used below, in a list of possible individual welfare metrics taking into account both consumption and leisure. The measures proposed and used in this paper are, therefore, not new. What is novel, is that the empirical rank correlations of welfare orderings based on these different measures, can now be interpreted as showing the sensitivity of welfare orderings to ethical choices about how to deal with preference heterogeneity. Moreover, the empirical nature of our paper complements results from similar exercises in Hodler (2009), Luttens and Ooghe (2007) or Schokkaert et al. (2004), where the application of a proposed normative analysis in societies with heterogeneous preferences is confined to numerical simulations in highly stylized settings, and to Lockwood and Weinzierl (2012) who explicitly model the relative importance of ability and preferences in the observed variation in earnings in an extension of the standard optimal tax framework.

In order to provide this empirical evidence, we use microdata from the Socio Economic Panel (SOEP) for married couples in Germany for the period 2002–2005. We retrieve individual and household-specific preference heterogeneity, by estimating a structural discrete choice model of female labour supply similar as e.g. in Aaberge et al. (1995) or Van Soest (1995). We use this preference information to construct welfare orderings of households according to different metrics of welfare, each embodying different ethical choices concerning the preference heterogeneity in the consumption-leisure space. We then move beyond the more descriptive analysis and discuss the different welfare implication of the welfare measures when analysing the 2007 reform of social security contributions, which lowered contributions for unemployment insurance from 6.5 to 4.2 %.

The rest of the paper is structured as follows. In Sect. 2, we briefly overview the problem of making interpersonal comparisons when preferences differ. We show how well-understood money metrics can help fix the dilemma between respecting preferences and making interpersonal comparisons. We focus on the normative interpretation of these metrics. In Sect. 3, we present the structural model of labour supply which informs us about preference heterogeneity. This information is then used in Sect. 4 to

calculate welfare metrics, compare the welfare orderings and discuss the sensitivity of the welfare impact of a policy reform with respect to the choice of welfare metric. Section 5 concludes.

2 The welfare metrics and their normative interpretation

2.1 Preference heterogeneity and welfare comparisons

Observed bundles of consumption and leisure result from individual choices, explained by means of preferences and constraints.[4] We define preferences in the (c, l)-space where c stands for consumption (or net income) and l for labour supply. Denoting the preference representation function by $u(c, l; \mathbf{z}_i)$, where vector $\mathbf{z}_i$ contains observable variables, partly explaining heterogeneity in preferences, the chosen bundle (c_i, l_i) by individual i is rationalized as a choice of his most preferred bundle, given his choice set:

$$(c_i, l_i) = \arg\max \left[u(c, l; \mathbf{z}_i) \,|\, c \leq f\left(I_i, w_i l; \mathbf{z}_i\right),\ l \leq 1 \right], \tag{1}$$

where $f(.)$ is a function representing the tax-benefit system, transforming non-labour income I_i and labour income $w_i l$, with w_i denoting the gross wage for individual i, into net income c.[5] In this framework, differences in outcomes for different individuals are explained by differences in preferences (vector $\mathbf{z}_i$), and differences in the budget constraint following from differences in gross wages (scalar w_i), differences in non-labour incomes (scalar I_i) and differences in the tax system related to vector $\mathbf{z}_i$. We illustrate a typical configuration for two individuals, denoted by subscripts a (Ann) and b (Bob) in Fig. 1, where for simplicity, we have assumed away the tax-benefit system.

Ann has a lower preference for leisure, in that, compared to Bob, she requires less compensation to work more hours. She also has a higher non-labour income than Bob, but a lower wage. The choices made by Ann and Bob are represented by bundles a and b, respectively. Ann works more and has a higher net income and less leisure. Bob works less, has more leisure, but a lower net income. The question at hand is: how to choose a metric $m(c_i, l_i; \mathbf{z}_i, w_i, I_i)$ which takes into account both preferences and constraints of individuals and allows to order individuals from worse to better off?

[4] In our empirical application, we only deal with welfare metrics at the individual level, since our application is restricted to labour supply choices of female spouses in couples with fixed labour supply of the husband. For a similar application for 11 European countries and the US, see Bargain et al. (2013). Conceptually, the welfare metrics can as well be applied to households, either in a unitary setting with one household preference ordering or with two individuals each described by its own preference ordering as is common in the collective household models (see Vermeulen 2002 for an overview, and Bloemen 2010 for a recent application and evaluation in the context of labour supply). For applications of the methodology of this section to aggregate analyses such as ranking countries by means of alternatives to GDP, see Fleurbaey and Gaulier (2009), Jones and Klenow (2010), and for an overview Fleurbaey (2009).

[5] In the empirical application, we will find that this deterministic part of the preferences (captured by observable vector $\mathbf{z}_i$) explains only part of the variation in choices for individuals facing the same constraints. The rest of the variation is due to 'unexplained heterogeneity'. At this stage, we do not elaborate the normative treatment of this unobserved heterogeneity. This means that we assume that two individuals with the same vector $\mathbf{z}$ do have the same preferences.

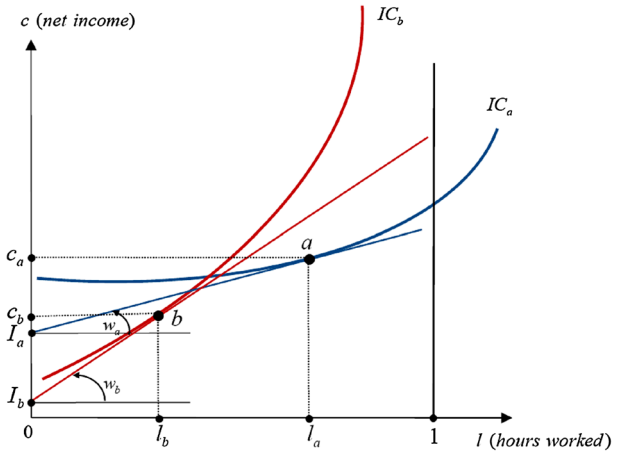

Fig. 1 The choice of Ann (a) and Bob (b) in the $(c, l)-$space with different preferences, different unearned incomes and different wages

That this is not an easy task has been well known for decades. The difficulty also formally appeared in the literature in the form of an incompatibility between two sets of axioms (see e.g. Fleurbaey and Trannoy 2003). On the one hand, there is Paretianity, requiring that, if all individuals weakly prefer social state x over social state y, the social ordering should also express a preference of x over y. In the following, we will refer to this intuition as "Respecting Preferences" or "Non-Paternalism". On the other hand, one needs axioms which embody some form of interpersonal comparability, such as Pigou–Dalton in multidimensional settings, or dominance of bundles.[6] With crossing indifference curves, one easily shows that both intuitive and attractive axioms clash and cannot be satisfied simultaneously.

The standard way out of this problem is to put aside the requirement of respecting preferences. To ensure consistency of welfare comparisons based on widely used concepts such as equivalent or compensating variations, one follows King (1983) and plugs the chosen bundles into a reference preference ordering, which is then used to calculate money metrics by using reference prices (which are the same for all individuals).[7] Normatively spoken, this boils down to impose some kind of objective criterion of welfare, which might be called a 'perfectionist' view of well-being. The analyst (or policy maker) introduces interpersonal comparability by fixing the welfare criterion independently from the preferences of the individuals. Otherwise stated, although preferences continue to play their full role in the determination of the chosen bundles, preference heterogeneity is de facto assumed away in the normative phase of the analysis, to wit, in the step where interpersonal comparisons are introduced.

Yet, the incompatibility suggests that there is also another possibility. A recent strand of the literature explores the possibility to give priority to Paretianity and to fully

[6] Bundle dominance states that an individual who has more of all commodities cannot be considered worse off than an individual who has less of everything.

[7] For applications in the context of labour supply and tax reform, see Aaberge et al. (2004) and Aaberge and Colombino (2013).

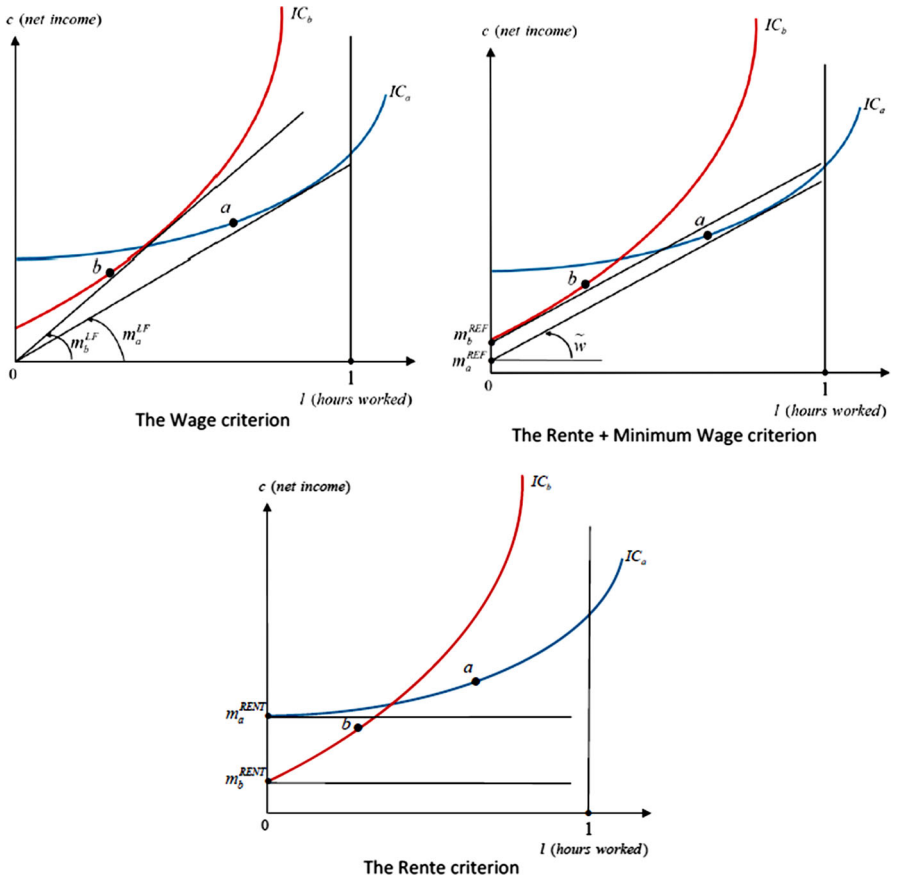

Fig. 2 The three welfare metrics

respect preference heterogeneity. The incompatibility result then inevitably points to the necessity of restricting the way one in which one implements interpersonal comparability. Recent proposals in Fleurbaey (2006, 2008) amount to restrict the interpersonal comparability by means of what is called *Subset Dominance*. Interpersonally comparable individual welfare levels are obtained by measuring individual welfare by means of nested sets, B_λ, where the set B_λ is implicitly defined by:

$$u(c_i, l_i; \mathbf{z}_i) = \max [u(c, l; \mathbf{z}_i) \,|(c, l) \in B_\lambda \,]. \tag{2}$$

The chosen bundle (c_i, l_i) on a given indifference curve is evaluated by indexing the curves by means of these equivalent sets, where $\lambda \leq \lambda'$ if and only if $B_\lambda \subseteq B_{\lambda'}$ and the situation of individual i is better the higher λ. Different metrics correspond to different specifications of the set B_λ in (2).

The first metric, illustrated in the top left panel of Fig. 2, is based on a specification of the equivalent set as:

$$B_{\lambda^{\mathrm{LF}}} = \left\{ (c, l) \,\middle|\, c \leq \lambda^{\mathrm{LF}} l, \quad l \leq 1 \right\}, \tag{3}$$

with a corresponding welfare metric for individual i equal to $m_i^{\mathrm{LF}} = \lambda^{\mathrm{LF}}(c_i, l_i)$, referred to as the 'Wage criterion'. [8] Chosen bundles a and b lead to welfare levels m_a^{LF} and m_b^{LF} by calculating the slope of the ray through the origin which delineates the subset of the (c, l)-space to which the indifference curve through the chosen point is tangent. In fact, this choice of the equivalent set amounts to the real wage criterion of Pencavel (1977), and the real wage metric W_5 in the list of Preston and Walker (1999).

The second class of examples rests on equivalent sets defined by

$$B_{\lambda^{\mathrm{REF}}} = \left\{ (c, l) \,\middle|\, c \leq \lambda^{\mathrm{REF}} + \widetilde{w} l, \quad l \leq 1 \right\}. \tag{4}$$

In this case, the indifference curves are indexed by means of equivalent sets which depend on a chosen reference net wage $\widetilde{w}$ and an unearned income λ^{REF}, where the corresponding individual welfare metric is then chosen to be this unearned income: $m_i^{\mathrm{REF}} = \lambda^{\mathrm{REF}}(c_i, l_i, \widetilde{w})$. Figure 2 illustrates the welfare metric for two choices of $\widetilde{w}$: a strictly positive $\widetilde{w}$: in the upper right panel and the special case of a reference net wage equal to zero in the bottom panel. This specific case of $m_i^{\mathrm{RENT}} = \lambda^{\mathrm{REF}}(c_i, l_i, \widetilde{w} = 0)$ is called the 'Rente criterion' by Fleurbaey (2006) and coincides with the intercept income of Preston and Walker (1999).[9]

2.2 Normative interpretation of the different metrics

As such, giving priority to respecting preferences, is not superior to the choice of giving up Paretianity and imposing one specific preference ordering. A convincing argument to choose for the subset dominance approach is given by Fleurbaey (2008) when countering the objection that the choice of reference prices and characteristics in the money metric utility approach is 'arbitrary':

> "if the equivalence approach depends on reference parameters, it can avoid arbitrariness if it develops an ethical theory of the choice of the reference. Some examples in the literature on fair social orderings show that rather natural axioms of fairness may force to adopt certain reference parameters". Fleurbaey (2008, p. 10).

[8] We use superscript LF to refer to the "Laissez Faire" description of this metric in Fleurbaey and Maniquet (2006).

[9] In this case, the equivalent set comes close to an implementation of interpersonal comparability in terms of reference bundles (as in Schokkaert et al. 2009). When the indifference curve is sloping upwards at $l = 0$, the tangency point of the equivalent set for a net wage equal to zero, becomes the corner solution. The Rente criterion, therefore, introduces interpersonal comparability by comparing individuals in the counterfactual situation 'as if they do not work', that is in terms of the reference bundle $(c, 0)$.

Otherwise stated, it might be easier to think about the ethical priors in terms of choosing these equivalent sets than in terms of e.g. a common utility function.[10] What are the implicit normative choices in the three individual metrics m_i^{LF}, m_i^{REF}, and m_i^{RENT}? First note that all three metrics fully respect preferences. That means that all metrics will increase when the individual moves to a bundle on a higher indifference curve of his or her *own preference ordering*, something which is not guaranteed when using a reference preference ordering. But preserving preference heterogeneity in the normative part of the analysis also confronts the analyst with a need to speak out normatively about how to deal with the fact that individual outcomes are the result of both preferences and constraints. This is precisely the topic of the literature on responsibility-sensitive egalitarianism, which addresses the issue by keeping individuals responsible for their preferences, but not for endowed circumstances. In order to operationalize this basic assumption, two competing interpretations evolved in the economic literature, namely the compensation and the (liberal) reward principle. The compensation principle states that inequalities due to endowed circumstances (i.e. not due to responsibility factors) should be removed. In contrast, the principle of liberal reward states that inequalities due to individual preferences are legitimate. Although similar at a first glance, both principles are logically independent and to some extent even in conflict with each other. Any individual welfare measure by which the analyst wants to make interpersonal comparisons therefore has to compromise on one of both principles. The individual welfare metrics presented in the previous subsection all give priority to the compensation principle, in that individuals with poorer (hypothetical) circumstances are always ranked worse off. But the measures embody different ethical priors on how to weigh people with different preferences differently e.g. by either favouring the industrious or the work averse.

Under the Laissez Faire criterion m^{LF} of Fig. 2 e.g. we judge two individuals as equally well off when they have the same hypothetical net wage rate, irrespective of the choices they make. Therefore, in this m^{LF} -measure, differences in preferences, leading to different choices, are considered not to be a sufficient reason for redistributing, or for ranking people as worse or better off.[11]

When choosing the Rente criterion, on the other hand, we offer maximal protection for people who have a larger distaste for working. With Bob's indifference curve cutting Ann's one from below in Fig. 2, we will always judge Bob to be worse off than Ann if they face the same constraint. From this perspective, choosing the Rente criterion as the welfare metric implements a normative choice of holding people with a strong aversion to work minimally responsible for these preferences.

[10] Fleurbaey (2005) gives the example of the metric designed to measure welfare in the multidimensional space of income and health. In that case, it seems natural (though not compelling) that one restricts interpersonal comparisons to the subset of the space where all individuals are healthy (instead of in bad health). And Schokkaert et al. (2009) argue that when constructing a measure of job satisfaction along the lines of subset dominance, one can better restrict interpersonal comparability to the subset of space where all individuals have a good job instead of when they have bad jobs.

[11] Framed in terms of a responsibility-compensation cut, one could say that this criterion holds people maximally responsible for differences in their tastes for leisure and is only willing to eventually compensate differences in (hypothetical) wage rates.

By moving away from the zero reference wage in the Rente criterion to the m^{REF}-metric with a strictly positive reference wage $\widetilde{w}$, it is easy to check graphically that for a given constellation of preferences (such as the ones of Ann and Bob in Fig. 1) the reference wage $\widetilde{w}$ in fact defines the subsets of metrics in the m^{REF}-set which will judge Ann to be better off than Bob (i.e. those metrics using a reference wage below $\widetilde{w}$) and the ones which will judge Ann to be worse off than Bob (i.e. those metrics using a reference wage higher than $\widetilde{w}$). Increasing the reference wage $\widetilde{w}$, therefore, is to be interpreted as changing the redistributive concerns. If we use the reference wage metric m^{REF}, we implicitly use social preferences in which we build in a redistributive bias in favour of distaste for work for all individuals with wages exceeding $\widetilde{w}$ (by ranking them lower), and against apparent laziness for all individuals with wages below $\widetilde{w}$ (by ranking them higher).

The empirical application on which we report in the next two sections is meant to answer the question how sensitive welfare orderings are with respect to the choice of the metric by means of which individuals are ordered, and hence to the normative choices made by the policy maker concerning preference heterogeneity. More precisely, we derive welfare orderings for the different measures derived above and show how sensitive the answer to the question "who are the poor? who are the rich?" is to the chosen metric.

3 Estimated preference heterogeneity

To apply the above metrics in a real-world context, we use German microdata from the SOEP for the years 2002–2005, which contains detailed information about the socioeconomic situation of households. The dataset is used as the input dataset for the Microsimulation model STSM (Steiner et al. 2008) which describes in detail the German tax-benefit system for the fiscal years 2001–2004. For a given gross wage, STSM allows to determine net income of the household for any chosen amount of labour supply. These detailed real-world budget constraints are combined with the observed choices of the individuals in the dataset to estimate a static structural labour supply model. Since the structural character of this labour supply model consists of a specification of the functional form of the preference representation function, this technique allows us to give empirical content to the preference heterogeneity of the previous sections.

Identification of structural models which relies only on cross-sectional variation in the data depends strongly on the chosen parametric functional form. Therefore, for the estimation, in addition to the variation across households, we also exploit variation in the tax-benefit system over the observed period 2001–2004. During this period the German Tax Reform 2000—the largest tax reform in post-war Germany—was implemented in several steps, and therefore, the income tax schedule in the observed years varies systematically. In particular, marginal tax rates were reduced and the tax exemption was increased such that all tax payers were affected by the reform. [12]

[12] For an overview, see Haan and Steiner (2005).

In the following, we describe the labour supply model and the functional specification chosen for the preferences; then, we give some information about the underlying data.

3.1 Specification of household preferences

We estimate household preferences by means of a static structural discrete choice model of labour supply, similar to Aaberge et al. (1995) or Van Soest (1995). The model is structural, because it starts from a specification of the utility function. And it is a discrete choice model because it reduces the choices of the individual (in this case the number of hours worked) to a finite number of discrete alternatives. The main advantage of this discrete specification over the continuous framework is the possibility to account for the nonlinearities in the budget set and to cope with the endogeneity of net household income in a relative straightforward way.

The discrete choice model starts from an empirical counterpart of the utility function in (1), by specifying the utility level of household i at a finite number of discrete chosen levels of labour supply. We index the discrete points by means of the subscript $j = 1, \ldots, J$. The state-specific level of utility of household i, denoted v_{ij}, at the $j = 1, \ldots, J$, discrete states consists of a deterministic and a stochastic part:

$$v_{ij} = u(c_{ij}, (1 - l_{ij}); \mathbf{z}_i) + \epsilon_{ij}, \tag{5}$$

where $u(c_{ij}, (1 - l_{ij}); \mathbf{z}_i)$ represents the deterministic part, and ϵ_{ij} is a stochastic random error term which varies independently between the individuals and the discrete points. Preference heterogeneity is captured by vector $\mathbf{z}_i$. As already mentioned in footnote 5 above, we will limit the analysis to observed preference heterogeneity and hence neglect household-specific heterogeneity which is unobserved. We assume that all unobservable effects are captured by the stochastic term ϵ_{ij}.

In this specific empirical application, we focus on the population of married households only. Moreover, we only consider the labour supply decision of the female spouse and assume that labour supply of husbands is exogenously determined.[13] That means that l_{ij} in (5) stands for female labour supply in household i (with $L_{ij} = 1 - l_{ij}$ denoting leisure time of the wife in household i), whereas c_{ij} refers to household net income. The latter consists of labour income of the wife and puts the exogenously determined labour income of the husband into non-labour income.

Similar to Aaberge et al. (2004) we use a Box-Cox functional form to specify the deterministic part of the utility function in (5):

$$u(c_{ij}, (1 - l_{ij}); \mathbf{z}_i) = \beta_c \frac{c_{ij}^{\alpha_c} - 1}{\alpha_c} + \beta_L(\mathbf{z}_i) \frac{(1 - l_{ij})^{\alpha_L} - 1}{\alpha_L}, \tag{6}$$

[13] We choose to focus on married couples since the economic literature e.g. Blundell and McCurdy (1999) has shown that behavioural labour supply responses of married women are particularly important.

where preference heterogeneity is introduced by means of taste shifters in the following form:

$$\beta_L(\mathbf{z}_i) = \beta_{L0} + \beta'_{L1}\mathbf{z}_i, \tag{7}$$

and vector $\mathbf{z}_i$ includes the age of both spouses, educational dummies, the number and age of children and a regional dummy. Preferences are determined by the parameters β_c, β_{L0}, β'_{L1}, α_c and α_L. The β-parameters determine the marginal utility of consumption and leisure, whereas the α-parameters determine the concavity of the utility function.

We are aware that the chosen functional form is relatively simple. But its main advantage is that it makes the calculation of the welfare metrics relatively straightforward. Moreover, more flexible specifications, in particular by inserting choice-specific intercepts, would certainly improve the model fit. But they come at the cost of a clear economic interpretation, which in terms of our aim to interpret preference heterogeneity in normative terms is a major disadvantage.[14]

The estimation procedure is based on the assumption that the error terms ϵ_{ij} are i.i.d. and follow an extreme value distribution. This gives an expression of the probability for each discrete working alternative, which results in the well-known conditional logit framework that can be estimated by maximum likelihood. We want to focus on the calculation of the welfare metrics and not on the most sophisticated labour supply model, as e.g. in Aaberge et al. (2004) or Blundell and Shephard (2012). Therefore, we make some simplifying assumptions in the estimation procedure. As already mentioned above, we do not account for unobserved heterogeneity. Haan (2006) has shown that unobserved heterogeneity does not significantly affect the labour supply elasticities when using a similar specification with cross-sectional data. Nor do we model potential restrictions on the labour market as in Aaberge et al. (2004) or Bargain et al. (2010). The findings of Bargain et al. (2010) imply that demand side constraints bias elasticities in particular for men and single women, but tend to be less severe for the labour supply decision of married women.

3.2 Data and descriptive statistics

SOEP is a representative household survey for Germany with sufficient socioeconomic information to derive the budget line of a household i.e. the net household income, and to estimate labour supply behaviour.[15] For this analysis, we use the data collected in 2002–2005, with income information about the tax years 2001–2004. We restrict the sample to married households with a wife aged between 20 and 60 who is not self-employed, retired or in full-time education. Moreover, we consider only households in which the husband is working full time i.e. more than 30 h per week. This gives us a sample of 9455 observations. For female labour supply, we define $J = 5$ discrete

[14] To guarantee positive first derivatives with respect to consumption and leisure, transformations of the coefficients might be necessary. But in our empirical application, the first derivatives were found to be positive for all households.

[15] For a detailed description of the SOEP, see Wagner et al. (2007).

working alternatives: non-participation, two part time alternatives, full-time work and overtime.[16]

To derive net household income according to the tax legislation in Germany for the observed years at each discrete alternative of working hours, we use the microsimulation model STSM (Steiner et al. 2008). The microsimulation model captures all relevant rules of the tax-benefit system including the changes following the Tax Reform 2000, mentioned above.[17] More precisely, we use the microsimulation model to calculate gross household earnings as the sum of observed earnings of the husband and the state-specific earnings of the wife for each discrete hours point. Gross earnings of the wife are simply the state-specific hours multiplied by her expected market wage. For working women, we take the observed wage information as their market wage, while for the non-working, we impute an expected market wage using an estimated wage equation with selection correction. The wage equation includes the relevant individual-specific information such as education and experience and is separately estimated for women in East and West Germany. As an exclusion restriction, we rely, as is common in the literature, on non-labour income of the wife, in particular the earnings of the husband, and on the number of children younger than school age. For a more detailed discussion of the wage estimation and a presentation of the empirical results for a slightly different sample, namely married women aged 20–55, see Haan (2010). [18] The information on gross earnings is the key input for the microsimulation model which describes, in detail, all relevant transfer programmes, social security contributions and income taxation and which delivers the state-specific net household income c_{ij}. Leisure time at each hours point is simply the time endowment $T = 80$ minus working time.

Table 1 shows the overall distribution of the households at the five alternatives. We also show average working hours and average monthly net household income and the shares by region, by education level and by the presence of children younger than 3 years old. The data reveal the relatively low labour market attachment of married women. About 30 % of all married women are not working, slightly over 40 % works part time and less than a quarter of all married women work regular hours or more. Since in our sample, all husbands work at least 30 h, the net household income distribution between the five discrete states is not very unequal. In addition, this is partly related to the joint taxation with full splitting which leads to high marginal tax rates for the secondary earner.

Table 1 shows interesting differences in the distribution across the employment states by region, education and family composition. In our sample, roughly 20 % of all households live in East Germany, but we only find 13 % East Germans amongst

[16] The median of the empirical distribution in the following intervals define the discrete points: 0, [0–15], [16–34], [35–40], > 40. The estimation results are robust to changes in the approximation of the distribution of working hours.

[17] In several papers, e.g. Bargain et al. (2010), the effect of the tax-benefit system on the working incentives are discussed in detail by analysing budget lines for different household types. Given the focus of this paper, which is on the normative analysis, we refer the reader to the previous studies for a detailed account of the translation of the German tax-benefit system into budget constraints.

[18] Estimation results for this wage equation can be obtained by the authors upon request.

Table 1 Discrete employment states

Employment status	Share in %	Working hours per week	Net income per month (€)	Share of households (%)		
				Living in East Germany	With low education	With children younger than 3
1 Not working	30.2	0	2,727	13.7	16.8	27.2
2 0–15 h	16.3	10	3,048	5.6	13.9	14.1
3 16–34 h	28.1	23	3,353	18.5	7.8	4.8
4 35–40 h	19.7	38	3,744	40.6	9.0	2.1
5 >40 h	5.8	42	3,876	49.5	3.5	2.9

The sample consists of 9,455 married households where the husband is working at least 30 h. The second column gives median working hours for the intervals 0, [0–15], [16–34], [35–40], > 40, and this median is used to define the discrete employment states. The share of East German households in the population is 20 %, 11.5 % of all women are low educated, i.e. 9 years of school or less, and 12.4 % of all households have a child younger 3 years. *Source*: SOEP, wave 2002–2005 and STSM

the non-working women, and even less amongst part time work. On the other hand, the share of East Germans in the subset of households where the wife is working full time is close to 40 %. For overtime work, the overrepresentation of East Germans is even larger. By education, we find that amongst non-working women, the share of low educated is above the average. The opposite holds for the family composition. Close to 30 % of non-working women have a child younger than three years, as opposed to only 3 % of those working full time or more hours.

3.3 Estimation results

Table 2 presents the estimated parameters of the Box-Cox utility function in (6).

Parameters α_c and α_L, both smaller than 1, indicate that the utility function is concave with respect to consumption and leisure time. For consumption, the curvature comes close to a logarithmic functional form (which would be the case if $\alpha_c = 0$), and the concavity is more pronounced for leisure. As expected, households value consumption positively ($\beta_c = 3.134$ being positive) and - on average - women also value leisure time positively ($\beta_{L0} = 0.799$). However, we find significant preference heterogeneity by observable characteristics. In line with previous studies, we find that the taste for leisure increases with the presence of children, in particular for children younger than 3 years. We find positive effects of the educational dummies, where the reference category is high education. This implies that *ceteris paribus* women with low and medium education have a higher preference for leisure than women with the highest educational degree. Finally, we find important differences between women in East and West Germany. In line with the descriptive statistics of Table 1, women in West Germany have a significantly lower inclination to work. This different pattern in female employment behaviour has often been analysed and is mainly explained by the different history and socialization of the two parts of Germany before the reunification.

Table 2 Estimated parameters of Box-Cox utility function

	Coefficient	Standard error
Preferences for consumption		
β_c	3.134	0.224
α_c	0.524	0.053
Preferences for leisure		
β_{L0}	0.799	0.152
β'_{L1} (taste shifter dummies)		
Age of wife	2.168	0.533
Age of husband	−0.638	0.472
Child younger than 3	2.267	0.205
Child between 4 and 6	1.052	0.106
East Germany	−0.891	0.087
Low education	0.354	0.074
Medium education	0.295	0.051
α_L	−1.527	0.134

α_c and α_L determine the concavity of the utility function with respect to consumption and leisure. β_c and β_L determine the marginal utility of consumption and leisure. Estimation includes time dummies. *Source*: SOEP, wave 2002–2005; Number of observations: 9,455.

Before we turn to a more detailed discussion of the preference heterogeneity in our estimated model, we provide evidence on the model fit in Table 3. As discussed above, our chosen functional form is fairly simple and therefore less flexible than in other studies e.g. Van Soest (1995). The advantage of this functional form, however, is a clear economic interpretation and the possibility to calculate the welfare metrics in a straightforward way. Moreover, despite the restriction in flexibility, the estimated parameters enable to reproduce fairly accurately all important features of the observed data presented in Table 1. This is demonstrated in Table 3 which shows observed and predicted (sub)-population shares at the different discrete states of labour supply. We predict the share of non-working women, of women working less than full time and of women working full time and more, quite precisely. Even more important for our application is that also the observed and predicted shares in the different sub-populations match quite well. Our estimated model replicates the increasing share of East German women by working hours, as well as the relatively higher share of women with children and of low-educated women in the non-working subgroup.

In Table 4, we present the preference heterogeneity by means of the variation in the marginal rates of substitution (MRS) for different subgroups. Given that for computational reasons, we derive the welfare metrics only for households observed in the year 2004, see next section for a more detailed explanation; we discuss the preference heterogeneity only for this population. The table includes the (sub)-population mean and the related standard deviations. For all households in the sample, we calculated the slope of the indifference curve at the same bundle of 40 h of weekly labour supply, and a net monthly income of 2,000 euros. Note, that for the calculation of the MRS, we only make use of the deterministic part of the utility function and neglect the stochastic part. In this respect, we consider the MRS only as a convenient summary of all estimated coefficients of the preferences. However, as discussed below, we will account

Table 3 Fit of the model: observed and predicted share of households in different employment states

Employment status		Subsample of households							
		Full sample		Living in East Germany		With low education		With children younger than 3	
		Obs.	Pred.	Obs.	Pred.	Obs.	Pred.	Obs.	pred.
1	Not working	30.2	26.3	13.7	11.0	16.8	14.6	27.2	24.7
2	0–15 h	16.3	27.0	5.6	14.0	13.9	13.8	14.1	16.5
3	16–34 h	28.1	21.4	18.5	18.5	7.8	11.3	4.8	7.1
4	35–40 h	19.7	13.6	40.6	37.2	9.0	6.8	2.1	1.0
5	>40h	5.8	11.7	49.5	47.3	3.5	5.1	2.9	0.4

The sample consists of 9,455 married households where the husband is working at least 30 h. The second column gives median working hours for the intervals 0, [0–15], [16–34], [35–40], > 40, and this median is used to define the discrete employment states. 'obs' indicates the observed share and 'pred' the share predicted by the model. The share of East German households in the population is 20 %, 11.5 % of all women are low educated, i.e. 9 years of school or less, and 12.4 % of all households have a child younger 3 years. *Source*: SOEP, wave 2002–2005 and STSM

for the stochastic part in the calculation of empirical magnitudes such as labour supply elasticities and for welfare metrics.

The results are striking. On average, the MRS in this bundle is 8 euros, but the variation is large. According to the estimated preferences, East German women are willing to work an additional hour for less than half the compensation asked by West German women (4 compared to 9.1). The presence of young children increases the distaste for work dramatically. The slope of the indifference curves for lower-educated people is steeper than for higher-educated ones, and contrary to what one would expect, the preference for work is not lower, but higher for females above 55.

At the bottom of Table 4, we also provide information about the size of the behavioural responses with respect to changes in financial incentives by simulating labour supply elasticities. In particular, we increase female gross wages by 1 %, and given the estimated parameters, we simulate relative changes in expected average participation rates and the relative change in expected weekly working hours. Expected values are calculated as weighted averages of the different employment states, where we use the conditional logit probabilities related to the stochastic element of the utility function for the J discrete labour market choices as weights. In addition to the point estimates, we also present bootstrapped confidence intervals.[19] The magnitude of the elasticities is very much in line with previous studies and suggests that women only modestly respond to changes in their budget line.

3.4 Empirical welfare metrics

To calculate the welfare metrics defined in Sect. 2 using the preferences estimated in this section, we simulated for each household in the sample expected labour sup-

[19] We use a parametric bootstrap with 1000 draws from the estimated variance-covariance matrix.

Table 4 Marginal rates of substitution for different groups

	Marginal Rate of Substitution in € per hour	Standard deviation
Whole sample	8.0	4.2
West German household	9.1	3.9
East German household	4.0	3.2
Children younger than 3	17.5	2.9
Children younger than 6	13.9	4.6
Low education	9.7	3.5
Medium education	9.1	3.5
High education	7.3	4.5
Female younger than 25	11.0	6.4
Female between 25 and 55	8.1	4.3
Female older than 55	6.9	2.2
Labour supply elasticities of 1 % increase in gross wages		
Change in participation rate (in %)	0.178	(0.166–0.186)
Change in working hours (in %)	0.439	(0.410–0.473)

Marginal rates of substitution were calculated in the bundle $(c, l) = (2000, 40)$. Labour supply elasticities were obtained by increasing female gross wages by 1 %. *Source*: SOEP wave 2005; Number of observations: 2,077

ply, expected disposable income and the expected welfare metric. For the latter, we followed the same approach as for the other expected values i.e. we first calculate the welfare metrics at each discrete labour market state, and then, in a second step, we calculate the expected values as a weighted average, using the probabilities for the J discrete choices as weights. Using expected values actually integrates out the random term in the welfare metrics. To calculate the welfare metrics at each of the discrete states, we used the analytical or numerical procedure described in the working paper version of this paper (Decoster and Haan 2010). Since the numerical procedures are computationally demanding, we decided to reduce the sample size to the 2,077 households which were observed in 2004 for the calculation of the welfare metrics.

4 Who are the poor? Who are the rich? Who are the gainers? Who are the losers?

We present the sensitivity of the welfare ordering to the chosen normative framework for individual welfare measurement in three stages. First, we compare the ordering of households from worst to best-off for each welfare metric in a stylised setting where we removed differences in budget constraints, and households only differ in their preferences. Next, we produce an analogous picture for our real-world sample of households, where differences in preferences interact with differences in gross wage rates and non-labour income. Finally, we also investigate the sensitivity of a distribution of gainers and losers of a tax reform for the chosen welfare metrics.

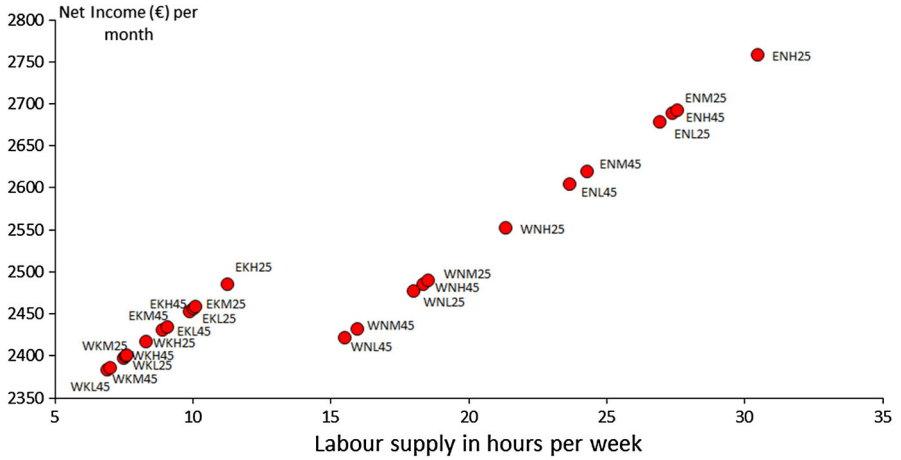

Fig. 3 Expected labour supply and net income for 24 stylized households

4.1 Results for 24 stylised households

We defined a set of stylised households by fixing the female gross wage at € 10 in a household where the husband is working full time (38 h a week) at a gross wage of € 15 per hour. The combination of two regional values (E for East and W for West German), the possibility that children younger than 3 are present (K if present, N if not), three levels of education (L for low, M for medium and H for high) and two selected ages (25 and 45) produce 24 typical households. With given gross female wage and a given non-labour income, these stylized households then only differ in two respects: (i) in their preferences and (ii) in whether they are eligible for child benefits.

Figure 3 shows the results of simulating labour supply for the females in these households, and the corresponding monthly net income. All results are in expected values. The preference heterogeneity induces large variations in labour supply behaviour, ranging from about 6 h a week to nearly 30 h a week. All households choose a bundle on the budget constraint, and the figure clearly reveals the upward shift of the budget constraint due to the presence of child allowances in the tax-benefit system. Besides the effect of young children, the figure mainly illustrates that females in Eastern German households, in general, work more than Western German ones. The whole northeast part of Fig. 3 is made up of East German households. Only if they received less education and are older (in this case 45 years old, see label ENL45), they reduce their labour supply.

The different choices in Fig. 3 obviously lead to different net incomes for the households. Apart from child allowances, working more also leads to a higher net income of the household, since all households have the same gross wage and the same non-labour income. Therefore, the young East German household with no kids and high education who works most (label ENH25) is considered to be the best-off in terms of net income, whereas the older West German household with kids and a middle

Table 5 Position in the welfare ordering of 24 stylized households

Household type	Labour supply hours/week	Net income €/ month	Position in welfare ordering based on			
			Net income	Rente criterion m^{RENT}	m^{REF} with reference wage $\widetilde{w} = €\ 7$	Wage criterion m^{LF}
WKL45	6.9	2,382	1	6	17	24
WKM45	7.0	2,385	2	7	19	23
WKL25	7.5	2,397	3	8	21	22
WKH45	7.6	2,399	4	9	22	21
WKM25	7.6	2,400	5	10	23	20
WKH25	8.3	2,416	6	11	24	19
WNL45	15.5	2,421	7	1	12	12
EKL45	8.9	2,430	8	12	20	18
WNM45	16.0	2,431	9	2	11	11
EKM45	9.1	2,434	10	13	18	17
EKL25	9.9	2,453	11	14	16	16
EKH45	10.0	2,456	12	15	15	15
EKM25	10.1	2,458	13	16	14	14
WNL25	18.0	2,477	14	3	10	10
EKH25	11.3	2,485	15	17	13	13
WNH45	18.4	2,485	16	4	9	9
WNM25	18.5	2,489	17	5	8	8
WNH25	21.3	2,552	18	18	7	7
ENL45	23.7	2,604	19	19	6	6
ENM45	24.3	2,619	20	20	5	5
ENL25	27.0	2,678	21	21	4	4
ENH45	27.4	2,688	22	22	3	3
ENM25	27.6	2,692	23	23	2	2
ENH25	30.5	2,758	24	24	1	1

The label of a household in the first column is composed of four characteristics, (ABCD) resp. indicating, West/East, Kids/No kids, Low, Medium or High education, and age of the female in the household.

education level (label WKL45) who supplies the lowest amount of labour is considered to be the worst-off in income terms. This is presented in Table 5, where we have ranked the 24 typical families in increasing order of net income. The different columns show the position in the welfare ordering of each household according to different welfare metrics, "1", indicating the worst-off household and "24" the best-off one.

The sensitivity of the answer to the obviously relevant policy question "who are the poor? who are the rich?" to the normative choices underlying the different welfare metrics is tremendous. Household WKL45 is the poorest in terms of income, but quickly moves up the ladder of the welfare distribution when leisure is taken into account. Moreover, its position heavily depends on how the policy maker or social analyst

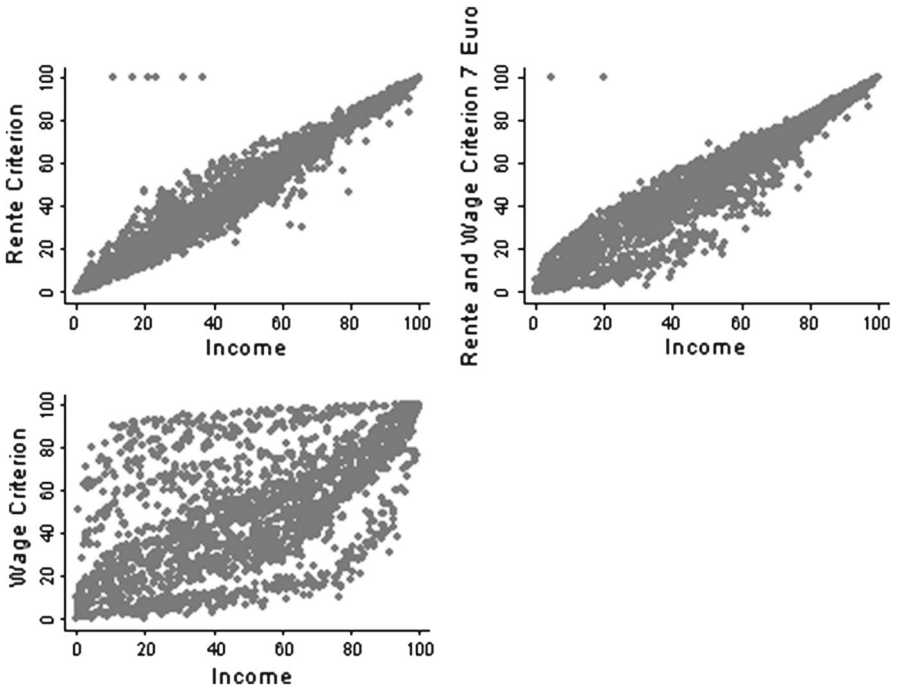

Fig. 4 Rank correlation between the ranking on the basis of net income and each of the three individual welfare metrics

weighs its preference characteristics relative to households who have preferences that are more favourable to supply labour. With the wage criterion e.g. which explicitly ignores differences in net incomes resulting from differences in preferences if gross wages are equal, the same household WKL45 ends up as the best-off household in the welfare distribution (rank 24 in the rightmost column of Table 5). The reverse holds for the household which is classified as best-off in net income terms (ENH25). With the wage criterion, this richest household is considered to be the worst-off (rank 1 on the bottom row of the rightmost column). These rerankings in the welfare ordering, based on clearly specified individual welfare metrics for this subset of households who only differ in their preferences, are striking. Preference heterogeneity not only matters in the positive analysis (to predict behaviour as precise as possible), it also matters in the normative phase of the analysis. Once the policy maker has chosen to respect preferences, he also has to make his weighing of differences in preferences explicit. Not unexpectedly, the degree to which he holds people responsible for their distaste for work dramatically determines the welfare ordering.

4.2 Welfare metrics for the population

The results of the previous subsection are exacerbated if, besides preference heterogeneity, we also introduce differences in gross wages and non-labour incomes. This is illustrated in Fig. 4 which compares the welfare orderings for the different welfare

Table 6 Composition of quintiles of the welfare ordering for different welfare metrics

Quintiles	Net income	Welfare ordering based on		
		Rente criterion m^{RENT}	m^{REF} with reference wage $\widetilde{w} = €7$	Wage criterion m^{LF}
Share of East German households (20 %)				
1	33.9	31.7	51.0	66.3
2	22.7	20.5	19.5	14.5
3	16.6	17.5	9.9	10.3
4	14.2	16.4	11.1	6.5
5	14.7	15.9	10.6	4.3
Share of households with low education (11 %)				
1	21.6	20.7	13.7	7.5
2	14.0	13.3	16.1	16.1
3	11.1	12.0	14.7	15.1
4	6.5	7.0	8.0	10.1
5	2.4	2.7	3.1	6.7
Share of hh's with children younger than 3 (11.5 %)				
1	14.9	11.8	3.6	0.0
2	14.9	14.9	13.7	1.4
3	9.9	11.1	13.5	2.9
4	9.4	10.4	15.7	13.0
5	8.7	9.6	11.3	40.5

This table presents the population shares of three demographic subgroups in the quintiles of different welfare orderings. Expected Welfare effects are derived from simulated labour supply behaviour. *Source*: SOEP, wave 2005; Number of observations: 2,077

metrics. More precisely, for each metric, we calculate the relative position of each household in the welfare ordering and compare the different rankings by means of a scatter plot. If all individuals are ranked in the same position for two metrics, the scatter is displayed as a diagonal one. We compare all measures with the net income criterion.

The upper left panel, with the comparison between the Rente criterion and the pure net income measure, shows that, not surprisingly, taking leisure into account clearly matters. Although there is some concentration on the diagonal, the orderings of the two measures clearly differ, but the introduction of variation in ethical priors about how to weigh differences in preferences is obviously even more important. The m^{REF}-criterion with a reference wage of € 7 still correlates quite well with the Rente criterion itself. Once we move to the Wage criterion m^{LF}, the correlation is weak or even non-existent.

The normative significance of this finding is further illustrated in Table 6. There we answer the same question "who are the poor?" and "who are the better off?" by describing the presence of households with certain characteristics in the different quintiles of the welfare distribution produced by a given metric. We consider three characteristics which are closely related to preference heterogeneity: living in East Germany, having young children and being lowly educated.

The results are striking when reading the table across the different columns. Take the first row, which shows the presence of East Germans in the bottom quintile of

the welfare distribution, and remember that about 20 % of the sample is living in East Germany. When the welfare ordering is based on disposable income alone, East Germans are clearly overrepresented in the poorest quintile. They do work more, but seemingly, their gross wages and their non-labour incomes are lower. Moving to the second column (the Rente criterion) is a move towards a criterion which also takes into account leisure. And yet, the harder working East Germans do not move down the welfare ranking because they work more. The reason is that, under the Rente criterion, they are pushed out of the bottom of the welfare distribution by those individuals who have a more pronounced distaste for working. The Rente criterion favours individuals with a distaste for work, ceteris paribus. Moving further to the right in the first row, across the columns of the table, shows how sharply the share of East Germans increases in the bottom quintile, when changing the ethical priors. When we hold individuals more responsible for their preferences w.r.t. the labour-leisure choice, and introduce a favourable treatment of the industrious relative to the work-averse individuals, the policy analyst will find that the bottom quintile of the welfare ordering is filled with 66.3 % East Germans, which is more than double of the share with the Rente criterion.

The same story holds for the other characteristics. The share of households with a lowly educated female in the bottom quintile, drops from 21.6 % under the Rente criterion to 7.5 % under the Laissez Faire criterion. And the 12.5 % of the bottom quintile which consists of households with children younger than three disappears completely from the bottom of the distribution. They appear to be predominantly well off (40.5 % of the top quintile) when the policy analyst considers their lower preference for work not as a legitimate reason for redistribution.

The interpretation of these striking changes in the composition of the quintiles of the distribution in Table 6 can, of course, be contaminated by correlation between the different characteristics. In Table 7, we therefore investigate whether the above findings are robust when we control for this correlation. We present results from multivariate regressions of the different welfare metrics on observed characteristics, viz. by region, education, presence of young children and non-labour income.[20]

The Rente criterion and the m^{REF}-criterion are defined in terms of monthly non-labour income. The Wage criterion m^{LF} is expressed in its monthly full-time equivalent. The coefficients can therefore be interpreted in monetary terms, although a direct comparison of the wage criterion with the other ones requires caution. Overall, the findings of Table 6 seem to be robust even after controlling for correlation between the characteristics. We find strong and significant differences in the welfare metrics by observed demographics which can be related to preference heterogeneity. *Ceteris paribus* net income is higher for women in East German households, lower for lowly educated females and lower for females with young children.[21] When the policy analyst moves to the Rente criterion, East German women (who have a lower preference for leisure) are judged to be even more better off than when using net income, and

[20] Note that for comparability, we always use expected rather than observed household income.

[21] The positive effect of the East German dummy on net income follows from the fact that we control for non-labour income (i.e. mainly the income of the husband), which is higher for West German households. A regression without this non-labour income as explanatory variable gives the expected negative sign for the East German dummy on net income.

Table 7 Regression of the different welfare metrics on demographic characteristics

	Net income	Welfare ordering based on		
		Rente criterion m^{RENT}	m^{REF} with reference wage $\widetilde{w} =$€7	Wage criterion m^{LF}
East Germany	118	191	−49	−192
	(23)	(18)	(19)	(10)
Low education	−169	−112	−92	−28
	(28)	(23)	(24)	(13)
Child younger 3	−212	−156	−41	385
	(30)	(24)	(25)	(14)
Child between 3 and 6	−243	−206	−123	111
	(27)	(22)	(23)	(12)
Age wife	3.5	1.0	4.6	7.2
	(2.4)	(1.9)	(2.0)	(1.1)
Age husband	5.0	5.0	4.1	1.4
	(2.3)	(1.9)	(1.9)	(1.1)
Non-labour income in (1,000)	456	452	473	209
	(8)	(7)	(7)	(4)
Constant	207	36	−65	44
	(63)	(51)	(53)	(29)

Coefficients are obtained by multivariate regressions of the welfare metric in monetary terms on demographic characteristics. Standard errors are between brackets. All welfare measures are expressed in Euros/1,000 per months. Expected Welfare effects are derived from simulated labour supply behaviour. *Source*: SOEP, wave 2005; Number of observations: 2,077

lowly educated females and females with young children (who have a strong preference for leisure) are considered less worse off, ceteris paribus, than when using the income criterion. The striking result however is that, even when we control for other observable characteristics, we do find rank reversals when switching to different metrics. East Germans e.g. are, ceteris paribus, considered worse off when using the reference wage, and also when using the wage criterion. This rightmost column of Table 7 suggests that, when measured by the wage criterion, welfare is about 192 Euros lower for East Germans, ceteris paribus, whereas they were considered to be 191 euros better off by means of the Rente criterion. The opposite holds for females in households with young children, and the welfare difference between the different measures is even larger. Ceteris paribus a household with young children is considered to be 212 euros worse off with the net income criterion, but are 385 euros better off with the wage criterion. We find these rank reversals for all characteristics. They are outspoken for the presence of children, but individuals with less education are no longer considered worse off neither, once the policy maker does no longer accept that preference characteristics, leading to a lower willingness to work, are a legitimate reason for redistribution.

Tables 6 and 7 not only illustrate the importance of taking leisure into account in the individual welfare measure. They also point to the importance of clearly specifying and founding the normative choices underlying redistributive activities in a setting where one respects preference heterogeneity.

4.3 Gainers and losers of a reform in work incentives

The previous section demonstrated how sensitive the welfare distribution is to normative principles in a setting which respects preference heterogeneity. However, in practice, policy makers are often more interested in identifying gainers and losers of policy reforms. To investigate whether the welfare *difference* is less sensitive to the chosen ethical priors, we evaluate a reform of the social security contributions, implemented in Germany in 2007. In particular, the contributions for unemployment insurance were reduced from 6.5 % of individual gross earnings to 4.2 %.[22] All employed with earnings above the exemption level for social security contributions were affected by this reform. Therefore, the induced labour supply incentives for married women are a priori unclear. On the one hand, there is an income effect related to the increase in net earnings of the husbands, but on the other hand, the direct effect for women induced a substitution effect which should increase labour supply.

Similar to the calculation of the labour supply elasticities, we used the labour supply model to determine the behavioural reaction of this reform in terms of the expected labour supply effects and calculate the expected welfare metrics before and after the reform as explained in Sect. 3.4.[23] First, note that it is in fact possible to rely only on ordinal utility information when describing the effects of a reform. This could be done e.g. by identifying winners and losers of the reform. However, to reveal the dependency on the normative choice of a welfare metric, we would also like to tabulate the number (or proportion) of winners and losers on groupings which depend on the chosen welfare metric. Unfortunately, since in the specific reform considered here, everybody wins (the reform being not revenue neutral), this kind of tabulation would not be informative at all. Therefore, in Table 8 we used the relative change in the individual welfare metric to rank the individuals in an increasing order of welfare gains. We partitioned the distribution of gains into quintiles and Table 8 describes the composition of these quintiles in terms of characteristics that were found to be related to preference heterogeneity.

The bottom quintile in Table 8 contains the households who have the smallest gain. The top quintile is populated by the households with the largest gains. According to the pure income measure which neglects leisure, East Germans are slightly underrepresented in the top quintile of gainers (18 % of this quintile consists of East Germans). The reason is to be found in the fact that, on average, West German wives benefit more

[22] The central idea of this reform was to reduce labour costs and therefore to improve the competitiveness of the German economy. At the same time, VAT was increased from 16 to 19 %. Given that we do not model consumption behaviour of individuals, we are unable to analyse the welfare effects of this second part of the reform.

[23] We find an increase in the expected female participation rate by 0.323 % in a 95 %-confidence interval of [0.29 - 0.357] and an increase in the expected working hours by 1.02 % in a 95 %-confidence interval of [0.95 - 1.108].

Table 8 Composition of quintiles of gainers and losers of a change in work incentives

Quintiles	Net income	Welfare ordering based on		
		Rente criterion m^{RENT}	m^{REF} with reference wage $\widetilde{w} = €7$	Wage criterion m^{LF}
Share of East German households (20 %)				
1	31.7	30.0	27.6	46.6
2	19.8	15.4	16.4	23.6
3	18.5	16.6	17.1	14.9
4	14.0	20.0	18.6	9.9
5	18.1	20.0	22.4	7.0
Share of households with low education (11 %)				
1	24.0	22.1	22.4	15.9
2	14.7	16.1	16.1	17.1
3	10.1	10.6	10.6	13.5
4	5.5	5.8	5.5	5.8
5	1.2	1.0	1.0	3.4
Share of hh's with children younger than 3 (11.5 %)				
1	23.8	20.2	23.3	8.2
2	13.0	14.7	14.2	7.7
3	9.4	7.9	9.4	9.1
4	9.6	13.3	10.1	10.6
5	1.9	1.7	0.7	22.2

This table presents the sub-population shares in the quintiles of gainers and losers of a reform based on different welfare orderings. We consider a reform consisting of a reduction in the contributions for unemployment insurance from 6.5 to 4.2 % of gross labour earnings. Expected Welfare effects are derived from simulated labour supply behaviour. *Source*: SOEP, wave 2005; Number of observations: 2,077

from the reform than married women in the East, because of the income effect related to the earnings of their husband. Sticking to this change in net income to identify winners and loser, we find that the quintile of (relative) losers of the reform is dominated by lowly educated people, and even more outspoken, by households with young children. These results are of course directly related to the labour market participation of these respective groups. The question is whether the identification of gainers and losers is robust with respect to choice of the individual welfare metric.

We therefore move to the right in Table 8 to use metrics which take up leisure (and the change therein) in the welfare metric, and fully account for preference heterogeneity between the individuals. The share of East Germans amongst the gainers of the reform remains close to 20 % when using the Rente criterion or the Reference Wage. However, when focussing on the wage criterion (m^{LF}), the share of East Germans is markedly reduced from 18 to 7 %. This illustrates the crucial role of the slope of the indifference curves (and hence the preference heterogeneity), not only in the calculation of the welfare level, but also for the welfare *difference*. A given net income change translates in a larger welfare gain (e.g. measured on the vertical axis at $l = 0$), the flatter the indifference curve is. With the Rente criterion e.g. one not only considers people with distaste for work as worse off in levels, one also considers that an increase in labour income is valued less by them.

The choice of metric also has an outspoken effect on where we classify the families with young children: the share in the lowest quintile (the relative losers) varies between 20.2 and 8.2 % when switching from the Rente to the Wage criterion. The difference is similar when focussing on the highest quintile. For education, the effect is especially striking in the quintile of households that incur the largest loss. Lowly educated households make up 24 % of the bottom quintile of the gainers distribution when using the income criterion. But this reduces to 15.9 % when taking leisure into account, and using the wage criterion.

5 Conclusion

Besides differences in budget sets, heterogeneity in preferences plays a crucial part in explanatory models of labour supply. But the incompatibility between respecting heterogeneous preferences and interpersonal comparability has confined applied welfare analysis to the case of comparability by means of a reference household or individual. Sensitivity analysis of the robustness of empirical results with respect to the choice of the reference household suggests that the choice of this reference preference is not very important (Aaberge et al. 2004).

Introducing a reference preference ordering is, however, only one way to escape the impossibility result. In this paper, we have followed a different route in the normative part of the analysis by calculating welfare metrics which fully respect preference heterogeneity but restrict the scope of interpersonal comparisons. We applied some of the measures developed in Fleurbaey (2006, 2008) and Fleurbaey and Maniquet (2011) and highlighted their different underlying normative priors in the empirical context of an estimated labour supply model. These by now standard discrete choice models of labour supply reveal considerable preference heterogeneity and hence are excellent candidates to illustrate the normative issues at hand. In this paper, we explored how this positive information could be fed into the proposed metrics and shed light on the empirical relevance of the choice to respect preference heterogeneity.

The results of the comparison of welfare orderings based on different metrics are striking. Not the inclusion of leisure into the welfare metric plays the decisive role, but the different normative treatment of the preference heterogeneity with respect to the labour-leisure choice. This indicates that the robustness of results with respect to the choice of the reference household in e.g. Aaberge et al. (2004) might have to do more with the removal of preference heterogeneity than with a robustness as such. The illustrative results have severe consequences for any policy advice which wants to incorporate distributional analyses against the background of preference heterogeneity (and respecting it). The answer to the question "who is worst-off" and "who is best-off" inevitably has to face the question how to treat people with different preferences differently. Does one consider preference characteristics as legitimate sources for compensation or not? If the answer is affirmative, one might go for a normative analysis based on, what is called in this paper, the Rente criterion. In that case, the difference between welfare ordering based on disposable income and a metric which includes leisure seems to be less important. If, however, one only considers differences in the budget constraints, as legitimate reasons for redistribution, one might opt for the wage

criterion. The correlation between the ordering based on disposable income and this wage criterion is very weak.

The purpose of this paper was not to discuss whether the observable preference characteristics used in this application are indeed legitimate sources of compensation and /or discrimination between individuals. Put even more generally, and as convincingly argued in Manski (2012), one can question whether at this stage our models and data are sufficiently rich to consider preferences derived from these models as sufficiently informative to evaluate policies. But we are convinced that this does not diminish the relevance of the illustration provided in this paper. Quite the contrary. The more progress we hopefully make in identifying and describing preference heterogeneity in the near future, the more urgent it becomes to properly integrate preference heterogeneity in the normative framework.

References

Aaberge, R., & Colombino, U. (2013). Using a Microeconometric Model of Household Labour Supply to Design Optimal Income Taxes. *Scandinavian Journal of Economics*, *115*(2), 449–475.

Aaberge, R., Colombino, U., & Strøm, S. (2004). Do More Equal Slices Shrink the Cake? *An Empirical Investigation of Tax-Transfer Reform Proposals in Italy, Journal of Population Economics*, *17*(4), 767–785.

Aaberge, R., Dagsvik, J., & Strøm, S. (1995). Labour supply responses and welfare effects of tax reforms. *Scandinavian Journal of Economics*, *97*, 635–659.

Auerbach, A. (1985) The Theory of Excess Burden and Optimal Taxation, in Auerbach, A. and Feldstein, M. (eds.) Handbook of Public Economics, Vol. 1, Elsevier Science Publishers, 61–127.

Bargain, O., Caliendo, M., Haan, P., & Orsini, K. (2010). 'Making Work Pay' in a rationed labour market. *Journal of Population Economics*, *23*(1), 323–351.

Bargain, O., Decoster, A., Dolls, M., Neumann, D., Peichl, A., & Siegloch, S. (2013). Welfare. *Labor Supply and Heterogeneous Preferences: Evidence for Europe and the US, Social Choice and Welfare*, *41*(4), 789–817.

Bloemen, H. (2010). An Empirical Model of Collective Household Labour Supply with Non-Participation. *Economic Journal*, *120*, 183–214.

Blundell, R. and McCurdy T. (1999), Labor Supply: a Review of Alternative Approaches, in: Ashenfelter O. and Card D. (eds.), Handbook of Labour Economics, Vol. 3A., Elsevier Science Publishers.

Blundell, R., & Shephard, A. (2012). Employment. *Hours of Work and the Optimal Taxation of Low Income Families, The Review of Economic Studies*, *79*(2), 481–510.

Boadway, R. (2012). Review of 'A Theory of Fairness and Social Welfare' by Marc Fleurbaey and François Maniquet. *Journal of Economic Literature*, *50*(2), 517–521.

Boadway, R., & Bruce, N. (1984). *Welfare Economics*. Oxford: Basil Blackwell.

Capéau, B., Decoster, A., De Swerdt, K. and Orsini, K. (2009), Welfare effects of alternative financing of social security. Calculations for Belgium, in: Zaidi, A., Harding, A. and Williamson, P., New Frontiers in Microsimulation Modelling, Chapter 17, 437–470, Ashgate.

Creedy, J., & Kalb, G. (2005). Discrete Hours Labour Supply Modelling: Specification, Estimation and Simulation. *Journal of Economic Surveys*, *19*(5), 697–734.

Creedy, J., & Hérault, N. (2011). *Decomposing Inequality and Social Welfare Changes: The Use of Alternative Welfare Metrics*.Department of Economics Research Paper Number: University of Melbourne. 1121.

Creedy, J., & Hérault, N. (2012). Welfare-improving income tax reforms: a microsimulation analysis. *Oxford Economic Papers*, *64*(1), 128–150.

Decoster, A. and Haan, P. (2010), Empirical welfare analysis in random utility models of labour supply, IZA Discussion Paper Number 5301.

Donaldson, D. (1992). On the aggregation of money measures of well-being in applied welfare economics. *Journal of Agricultural and Resource Economics*, *17*, 88–102.

Eissa, N., Kleven, H., & Kreiner, C. (2008). Evaluation of four tax reforms in the United States: Labor supply and welfare effects for single mothers. *Journal of Public Economics*, *92*(3–4), 795–816.
Fleurbaey, M. (2005). Health, and Fairness. *Journal of Public Economic Theory*, *7*(2), 253–284.
Fleurbaey, M. (2006). Social welfare, priority to the worst-off and the dimensions of individual well-being. In F. Farina & E. Savaglio (Eds.), *Inequality and economic integration*. London: Routledge.
Fleurbaey, M. (2008a), Fairness, Responsibility, and Welfare, Oxford University Press.
Fleurbaey, M. (2008b) Willingness-to-pay and the equivalence approach, Oxford Poverty & Human Development Initiative, OPHI Working Paper No. 25.
Fleurbaey, M. (2009). Beyond GDP: The Quest for a Measure of Social Welfare. *Journal of Economic Literature*, *47*(4), 1029–75.
Fleurbaey, M., & Gaulier, G. (2009). International Comparisons of Living Standards by Equivalent Incomes. *Scandinavian Journal of Economics*, *111*(3), 597–624.
Fleurbaey, M., & Maniquet, F. (2006). Fair Income Tax. *Review of Economic Studies*, *73*(1), 55–83.
Fleurbaey, M., & Maniquet, F. (2011). *A theory of fairness and social welfare*. Cambridge University Press: Econometric Society Monographs.
Fleurbaey, M., & Trannoy, A. (2003). The Impossibility of a Paretian Egalitarian. *Social Choice and Welfare*, *21*, 243–263.
Haan, P. (2006).Much ado about nothing: conditional logit vs. random coefficient models for estimating labour supply elasticities. *Applied Economics Letters*, *13*, 251–256.
Haan, P. (2010). *A Multi-State Model of State Dependence in Labor Supply, Labour Economics*, *17*, 323–335.
Haan, P., & Steiner, V. (2005). Distributional Effects of the German Tax Reform 2000—A behavioral microsimulation analysis, Schmollers Jahrbuch. *Journal of Applied Social Science Studies*, *125*, 39–49.
Hodler, R. (2009). Redistribution and Inequality in a Heterogeneous Society. *Economica*, *76*(304), 704–718.
Jones, C. and Klenow P. (2010), Beyond GDP? Welfare across Countries and Time, NBER Working Paper No. 16352.
King, M. (1983). Welfare analysis of tax reforms using household data. *Journal of Public Economics*, *21*, 183–214.
Lockwood, B. and Weinzierl, M., (2012), De Gustibus non est Taxandum: Theory and Evidence on Preference Heterogeneity and Redistribution, NBER Working Paper No. 17784.
Luttens, R., & Ooghe, E. (2007). Is it Fair to Make Work Pay? *Economica*, *74*(296), 599–626.
Manski, C. (2012), Identification of Preferences and Evaluation of Income Tax Policy, NBER Working Paper No. 17755.
Pencavel, J. (1977). Constant-Utility Index Numbers of Real Wages. *American Economic Review*, *67*(1), 91–100.
Preston, I., & Walker, I. (1999). Welfare Measurement in Labour Supply Models with Nonlinear Budget Constraints. *Journal of Population Economics*, *12*, 343–361.
Schokkaert, E., & Van de gaer, D., Vandenbroucke, F. and Luttens, R. , (2004). Responsibility sensitive egalitarianism and optimal linear income taxation. *Mathematical Social Sciences*, *48*(2), 151–182.
Schokkaert, E., Van Ootegem, L and Verhofstad, E. (2009), Measuring job quality and job satisfaction, FEB working paper 2009/620.
Steiner, V., Wrohlich, K., Haan, P., and Geyer J. (2008), Documentation of the Tax-Benefit Microsimulation Model STSM: Version 2008, Data Documentation 31, DIW Berlin.
Van Soest, A. (1995). Structural Models of Family Labor Supply: A Discrete Choice Approach. *Journal of Human Resources*, *30*, 63–88.
Vermeulen, F. (2002). Collective household models: principles and main results. *Journal of Economic Surveys*, *16*(4), 533–64.
Wagner, G., Frick, J., & Schupp, J. (2007). The German Socio-Economic Panel Study (SOEP) - Scope, Evolution and Enhancements, Schmollers Jahrbuch. *Journal of Applied Social Science Studies*, *127*, 129–169.